Rescuing concatenation with maximum likelihood using supermatrix rooted triples

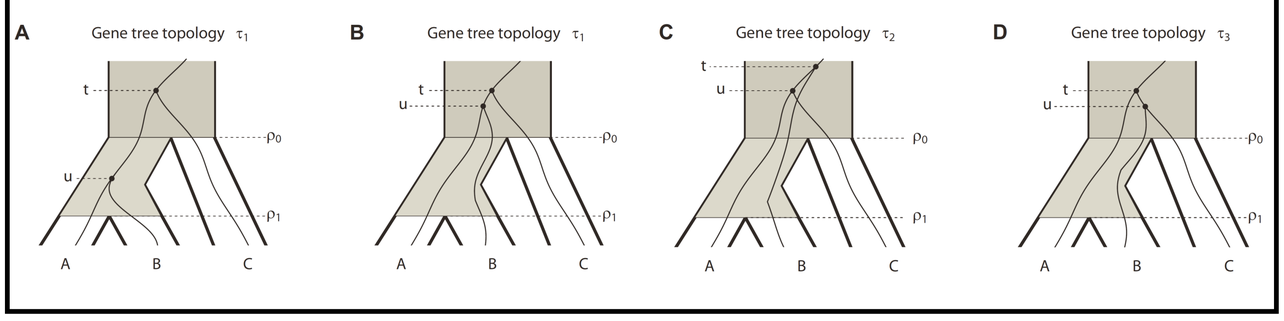
Michael DeGiorgio* and James H. Degnan†

*Center for Computational Medicine and Bioinformatics, University of Michigan
†Department of Mathematics and Statistics, University of Canterbury

Abstract

Concatenated alignments are often used to infer species-level reslationships. Previous studies have shown that analysis of concatenated alignments using maximum likelihood (ML) can produce misleading results. We develop a polynomial-time method that constructs a species tree through inferred rooted triples from concatenated alignments. We call this method SuperMatrix Rooted Triple (SMRT). We show that SMRT performs well in simulations and then show that it is a statistically consistent estimator of a clocklike species tree under a binary substitution model as well as other assumptions. SMRT is therefore a computationally efficient and statistically consistent estimator of species trees.

Gene trees in a model three-taxon species tree



SuperMatrix Rooted Triple (SMRT)

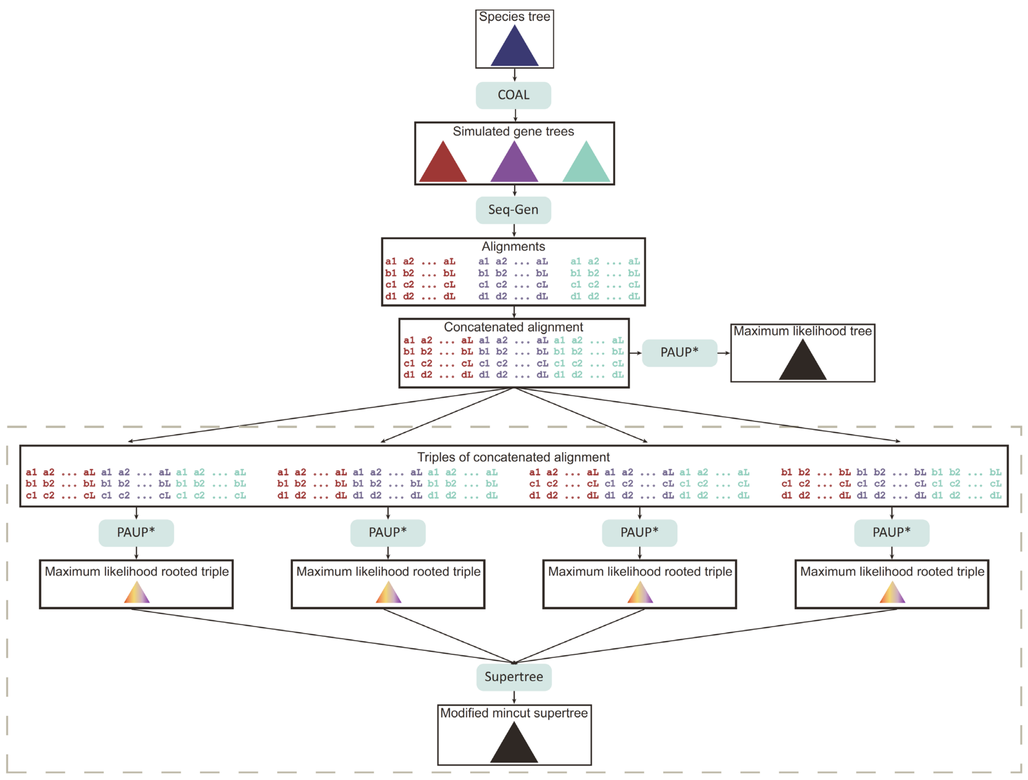
- New method for inferring species trees
- Concatenate m n-taxon alignments to create a supermatrix of n taxa
- Break supermatrix into n choose 3 supermatrices of three taxa
- Infer a rooted three-taxon tree (rooted triple) from each three-taxon supermatrix
- Use a supertree method to combine all rooted triples into an n-taxon species tree

Probability distribution of site patterns

- The probability of a site pattern $\mathbf{x}$ given species tree σ is obtained by integrating the probability of $\mathbf{x}$ given the gene tree topology τ_i and coalescence times t and u

$$P_\sigma(\mathbf{x}) = \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} P_\sigma(\mathbf{x} \mid t, u, \tau_1) g_\sigma(t, u, \tau_1)\, du\, dt + \sum_{i=1}^{3} \int_{\rho_0}^{\infty} \int_{\rho_0}^{t} P_\sigma(\mathbf{x} \mid t, u, \tau_i) f_\sigma(t, u, \tau_i)\, du\, dt$$

Simulation procedure



Statistical consistency

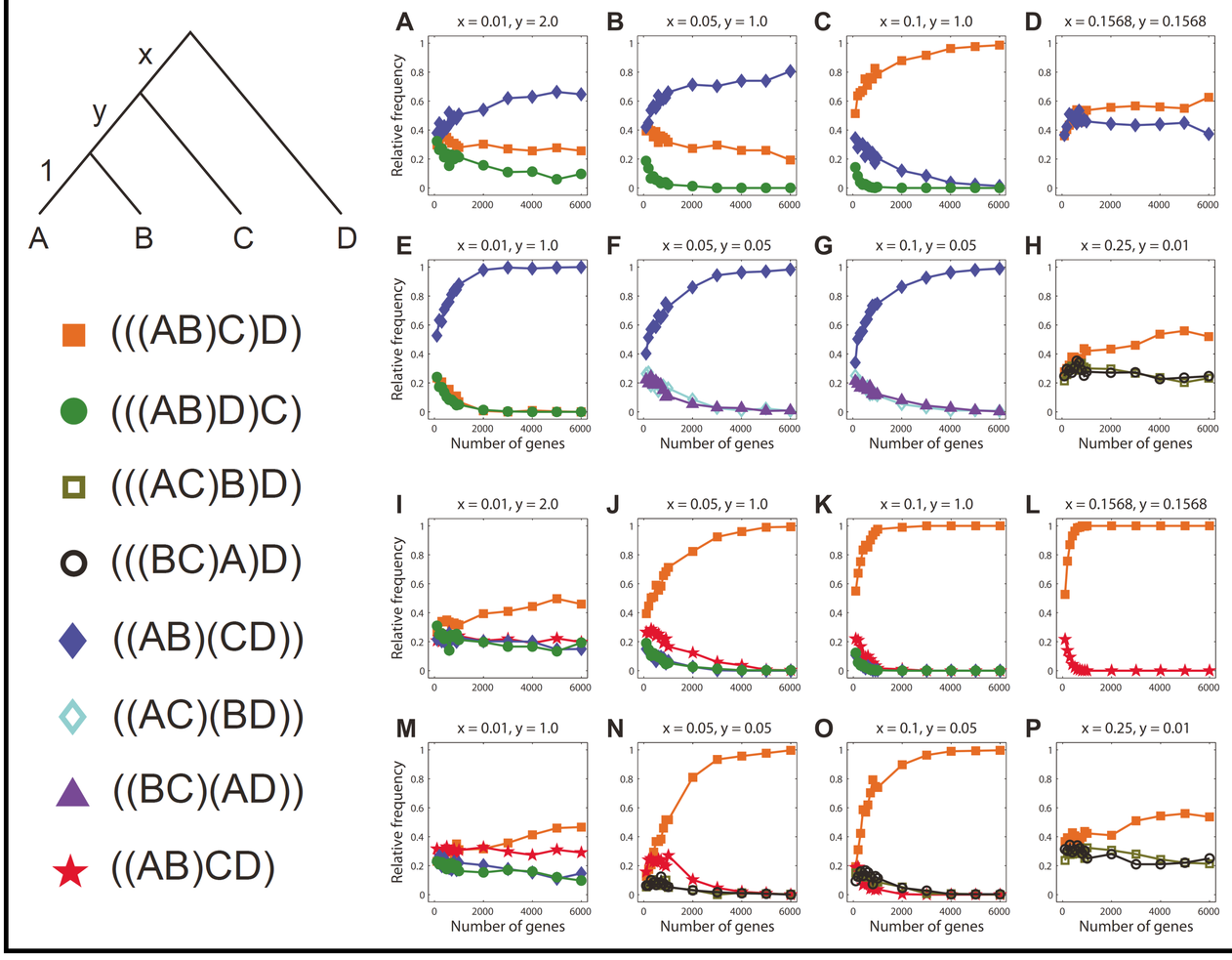
Assumptions

- Mutations occur under a binary substitution model
- The species tree is clocklike
- Incomplete lineage sorting is the source of discordance between gene trees and species trees
- There is no hybridization, horizontal gene transfer, or other gene flow between species
- There is no population subdivision within species
- Concatenated alignment does no grow “too quickly”
- A supertree algorithm is used with the property that if the input trees are compatible, then the supertree is a rooted phylogenetic tree which displays all input trees

Results

- **Lemma 1.** The proportion of sites with pattern $\mathbf{x}$ converges in probability to $P(\mathbf{x})$
- **Lemma 2.** SM-ML is a statistically consistent estimator of a three-taxon clocklike species tree
- **Theorem 3.** SMRT-ML is a statistically consistent estimator of a clocklike species tree with three or more taxa

Simulation results (SM-ML: A-H, SMRT-ML: I-P)



Conclusions

- Through simulations assuming a Jukes-Cantor substitution model and trees inferred using ML, SMRT-ML performs well compared to ML applied to a concatenated alignment (SM-ML)
- Through theory, we show that given certain assumptions, SM-ML is a statistically consistent estimator of a three-taxon species tree and therefore SMRT-ML is a statistically consistent estimator of a species tree on n taxa
- Only three tree topologies need to be investigated for each of the $n(n-1)(n-2)/6$ sets of three taxa. Using a polynomial-time supertree algorithm, SMRT can infer a species tree in polynomial time

Future directions

- How does SMRT perform when sampling multiple individuals per species?
- Is SMRT statistically consistent under more complex substitution models?
- Is SMRT statistically consistent using other methods such as distance and parsimony methods to infer the rooted triplers?